Evaluating a department-wide initiative incorporating computational methods into the undergraduate physics curriculum

Greg Gallagher,[1] Gautam Vemuri,[1] Danka Maric,[2] and Andrew Gavrin[1,*]
[1]*Department of Physics, Indiana University Indianapolis, 402 North Blackford Street, Indianapolis, Indiana 46202, USA*
[2]*STEM Education Innovation and Research Institute, Indiana University Indianapolis, 755 West Michigan Street, Indianapolis, Indiana 46202, USA*

(Received 19 December 2025; accepted 30 March 2026; published 28 April 2026)

We report on a department-wide initiative, implemented beginning in 2018, aimed at making computational methods a routine and integrated part of the undergraduate physics curriculum. Using survey data collected from undergraduate physics students across multiple years, we examined how students' attitudes toward computation and their self-efficacy with computational skills change as they progress through the program. The survey includes attitudinal items and self-efficacy ratings for a range of computational tools and methods, as well as retrospective assessments of students' initial skill levels. An exploratory factor analysis was conducted for both parts of the survey, and each part produced reliable, two-factor structures. Our analyses of student attitudes and self-efficacy throughout the program show that, at the program level, students generally report more positive views of computation and higher computational self-efficacy at later stages. However, these changes are not evenly distributed across the curriculum, with larger shifts concentrated in particular semesters. Together, these results show that our department-wide effort to integrate computational methods into the undergraduate curriculum supports growth in student attitudes and self-efficacy, but that this growth may be driven by specific experiences in the program rather than by progression alone. We also find that conclusions about growth in computational self-efficacy depend on how growth is measured. For some skills, growth based on present self-efficacy closely matches growth estimates that incorporate retrospective initial ratings, while for others, the two approaches differ substantially. Importantly, these differences are not uniform across skills but align with the underlying factor structure of the survey. Skills associated with specific computational tools or languages show relatively little response-shift effect, whereas skills associated with applied computational methods exhibit larger differences in growth estimates when retrospective measures are included. These findings suggest that retrospective pretest survey designs may be particularly important for assessing growth in methods-based computational skills. On the whole, this study highlights the importance of considering both curricular structure and measurement approach when evaluating our curriculum-wide initiative to integrate computation into undergraduate physics programs.

DOI: 10.1103/2zmq-2t62

I. INTRODUCTION

A. The increasing role of computation in physics education

Computational methods have become an increasingly prominent focus in physics education research (PER), with many programs experimenting with different approaches for incorporating computation into the undergraduate physics curriculum [1–4]. The visibility of this effort has grown in recent years, in part due to a report by the American Association of Physics Teachers (AAPT) offering strategies for integrating computation into undergraduate physics programs (e.g., offering dedicated courses in computational physics) [5]. A joint task force of the American Physical Society and the AAPT reinforced many of these recommendations and emphasized the importance of computational fluency for undergraduate physics students entering graduate school or the workforce [6]. From this joint task force emerged the Effective Practices for Physics Programs (EP3) initiative that provides guidance on best practices based on systematic reviews of current literature in PER [7]. Additional initiatives, such as the Partnership for Integration of Computation into Undergraduate Physics (PICUP)—which serves both as a professional community and resource

*Contact author: agavrin@iu.edu

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

platform—have further elevated the visibility of these efforts and underscored the importance of incorporating computation into the undergraduate physics curriculum [8].

Unfortunately, despite these calls to action, there remain relatively few reports of department-wide initiatives to integrate computation into the undergraduate curriculum. Notable exceptions to this are the physics programs at Oregon State University [9] and Lawrence University [10], where both programs have incorporated computational methods holistically throughout their undergraduate programs. Compounding this issue is the current lack of well-established instruments for evaluating the efficacy of these integration efforts. Neither the PICUP project website [8] nor PhysPort [11] offer suitable evaluation instruments, although some work has begun in this area [12,13].

B. Institutional response: A comprehensive curriculum initiative

This study draws on survey data collected from undergraduate students at an urban, public university with an R1 Carnegie Classification ("Very High Research Spending and Doctorate Production"). The Physics Department includes 15 full-time faculty and offers B.S., M.S., and Ph.D. degrees, with an undergraduate curriculum that follows a traditional structure comprising a two-course introductory sequence, six advanced courses, two upper-level labs, and a research capstone. For reference, Table I lists the required undergraduate courses in the order recommended by the department. An important factor to note which can introduce noise into the dataset is that some students will routinely take courses out of order.

Beginning in 2018, the Physics Department adopted a department-wide initiative to incorporate computational methods throughout the undergraduate curriculum. The overall goal of this initiative is to "make computation normal" throughout undergraduate physics coursework. In other words, the department aims for graduating students to become comfortable using computational tools and to view them as a routine component of a physicist's work.

To support this goal, the Physics Department established an expectation that at least 25% of student work in each undergraduate course include computational methods. Faculty are granted the flexibility to determine how this is implemented in their courses. Many instructors incorporate computation by structuring assignments such that 25% of the work involves computational tools, while others adopt semester-long computational projects that account for at least 25% of a student's grade.

C. Development of a novel survey instrument

Due to the lack of a suitable evaluation instrument in the literature, we developed the survey instrument described in this paper to evaluate our department-wide efforts to normalize computation [14]. Several considerations shaped the structure and dissemination of the instrument. In PER and more broadly in education research as a whole, the efficacy of an instructional intervention is often examined using a pretest/post-test design: students complete a pretest, experience an intervention, and then complete a post-test. This approach has deep roots in PER, including notably the seminal work of Halloun and Hestenes, whose early findings on conceptual learning in physics ultimately informed the development of the Force Concept Inventory [15,16]. Numerous PER synthesis studies have further documented the widespread use of pretest/post-test methods in PER [17–19]. In our case, however, the survey was developed *after* the department-wide initiative began, leaving no opportunity to collect baseline data before computation was introduced into the undergraduate curriculum.

TABLE I. A list of the currently required physics courses and associated course numbers for undergraduate physics majors. Courses are listed in the order recommended by the Physics Department.

Semester (year)	Course name	Course number
Spring (year 1)	Mechanics	152
Fall (year 2)	Heat, Electricity, and Optics	251
	Introduction to Computational Physics	299
Spring (year 2)	Introduction to Elementary Mathematical Physics	300
	Modern Physics	342
Fall (year 3)	Intermediate Mechanics	310
Spring (year 3)	Intermediate Electricity & Magnetism	330
	Electronics Laboratory	353
Fall (year 4)	Physical Optics	400
	Physical Optics Lab	401
	Quantum Mechanics	442
Spring (year 4)	Thermal and Statistical Physics	418
	Capstone Experience	490

Further, we were also concerned that students "don't know what they don't know," which complicates traditional pretest/post-test interpretations. A fundamental assumption of pretest/post-test designs is that a participant's understanding of the measurement scale (i.e., their frame of reference) does not change with repeated administrations of the instrument [20,21]. This assumption becomes particularly problematic in education contexts, whereby the goal of an intervention is usually to shift students' understanding of a domain [21,22]. Howard and Dailey [21] termed this issue *response-shift bias*, where changes in students' internal standard of measurement distort traditional pre-post comparisons. This phenomenon is conceptually related to the well-known Dunning-Kruger effect, in which individuals tend to overestimate their competence because they lack awareness of their actual knowledge level [23]. Because our initiative is to "make computation normal" for students, we fully expect their understanding of the computational domain to shift as they gain experience. This shift undermines the fundamental assumption of traditional pre-post designs, indicating that such an approach is unsuitable for our context.

To address the challenge of response-shift bias, researchers have developed and used *retrospective pretest* designs (also referred to as *reflective pretests*), in which participants, at the end of an intervention, assess both their current and *prior* understanding from the same frame of Ref. [24]. While this approach has gained attention in PER and related fields as a way to reduce response-shift bias in evaluations of student learning, retrospective and reflective pretests are most commonly applied to single-course or single-intervention studies [25,26]. Here, we seek to extend this methodology by designing a survey instrument that incorporates retrospective pretest elements and is administered repeatedly across multiple semesters and stages in the undergraduate program. This design allows us to examine response-shift effects in greater detail than typical single-intervention studies by applying retrospective pretest elements across the full undergraduate curriculum.

In its current form, the survey contains two main parts designed to capture complementary dimensions of undergraduate students' experiences with computation in the physics program. Part 1 asks students to rate their agreement (via a five-point Likert scale) with statements about the role of computation in physics and broader self-efficacy in using computational methods. Part 2 employs a retrospective pretest design by asking students to rate both their *initial* and *present* abilities via a ten-point scale, which allows us to account for response-shift bias among students.

Because our goal is to evaluate how students' attitudes and self-efficacy develop throughout the undergraduate curriculum, we administer the survey at the end of every semester rather than only once at graduation. This repeated administration allows us to identify when meaningful changes occur in student attitudes and self-efficacy and to examine course-specific and level-specific (i.e., 100-, 200-, 300-, 400-level) patterns in students' experiences with computation.

D. Present study and goals

We have previously reported on several components of this project, including the initial development of the survey instrument, early efforts to establish validity and reliability of the survey instrument, an exploratory factor analysis of the original part 1 items, and a preliminary analysis of growth in computational self-efficacy using items in part 2 of the survey [14,27,28]. These studies established both the foundation of the survey instrument and the motivating factors for refinement, including the revision of three part 1 items in 2024 based on the results of the factor analysis. Because these revisions changed the wording in those three items, we cannot directly compare these items to earlier administrations of the survey. Additionally, only two semesters of data exist for the revised items, which is not sufficient for the longitudinal, multilevel analyses conducted in this study. For these reasons, we have chosen to omit part 1 data collected after the Spring of 2024.

Our purpose is to extend prior work evaluating the department-wide initiative of incorporating computation into the physics curriculum, using data collected from our survey instrument since its introduction in 2018. To that end, we introduce the following specific research questions:

- RQ1: To what extent do students' self-efficacy and attitudes toward computational methods in physics change as they progress through the undergraduate physics program?
- RQ2: Do some courses contribute more than others to gains in students' self-efficacy and attitudes toward computation in physics, and if so, which courses contribute the most?
- RQ3: Compared to a typical pre-post survey design, does the inclusion of asking students to rate their *initial* and *present* skill level in part 2 of the survey (i.e., introducing a retrospective pretest component) introduce a quantifiable difference in our measurement of students' self-efficacy?

These questions frame the analyses presented in this study. Section II outlines our methodology for conducting the survey and each analysis. Section III presents the results for each analysis conducted, and Secs. IV and V provide discussion, conclusions, and future directions for continued refinement and administration of the survey instrument.

II. METHODS

A. Survey instrument overview

The survey instrument evaluated in this study consists of a background section followed by parts 1 and 2. The background section asks students to provide their name,

TABLE II. The survey items and convenient identifiers for part 1 of the survey. Students are asked to rate their agreement with each item on a Likert scale that is then converted to a numeric 1–5 scale for analysis.

Label	Survey item
L1	Computational, experimental, and analytical ("pencil and paper math") methods are all necessary in the field of physics.
L2	Using computational methods helps me understand physics topics.
L3	I can sometimes use computational methods to understand problems that I am unable to do analytically.
L4	I can judge whether a given problem is most easily or better solved by computational vs analytical means.
L5	Computational methods can be useful in conjunction with analytical methods for understanding physical phenomena.
L6	Using analytical and experimental methods helps me understand physics topics.
L7	I'm confident in the results of codes I develop.
L8	I feel well prepared for using computational methods in graduate school or a future job.
L9	I have used computational methods outside my classes, such as in a research project, internship, or job.

student ID number, and the physics courses in which they are currently enrolled. Part 1 includes nine attitudinal and general self-efficacy items in which students rate their agreement with statements about computation in physics using a Likert scale ranging from "strongly disagree" to "strongly agree," which is converted to a 1–5 numeric scale for analysis. Part 2 is divided into two parallel subparts, 2a and 2b, each containing the same set of 11 self-efficacy items (resulting in a total of 22 items for part 2) targeting a variety of computational platforms and skills. In part 2a, students rate their *present* skill level for each item on a 1–10 numeric scale. In part 2b, students again rank their skill for each item, but this time with respect to their *initial* skill level. The wording for part 2a is "Please rank your *present* ability on these topics/skills," and for part 2b is "Please rank your ability on these skills/topics *at the time you started taking physics at [this institution]*." In both cases, the emphasis is as it appears to the students. Table II lists the items and convenient labels for part 1 of the survey, while Table III lists the items and labels for part 2. These labels will be used throughout the remainder of this paper.

TABLE III. The survey items and convenient labels for part 2. These items are identical for parts 2a and 2b. Students are asked to provide their *present* and *initial* skill levels for each item, where "initial" refers to the skill level students had when they began taking physics in the program.

Label	Survey item
Q1	Excel (or other spreadsheets)
Q2	LaTeX
Q3	MATLAB
Q4	Other programming language
Q5	Rescaling a problem to eliminate units
Q6	Testing code to determine its overall correctness
Q7	Visualization of functions and other results
Q8	Matrix operations
Q9	Solving differential equations
Q10	Numerical integration and differentiation
Q11	Data analysis

Because part 1 of the survey measures attitudes and opinions using a Likert scale (via a more traditional pre-post format), while part 2 measures self-efficacy using a numeric scale (via a retrospective pretest structure), we treat the analysis of each section independently. Part 1 has undergone two substantive revisions since the survey was first administered in Fall 2018, whereas part 2 has remained largely stable. The only modification to part 2 of the survey involved reordering the items so that students rated their *present* skill before their *initial* skill. The content of the items in part 2, however, has remained unchanged, allowing us to use the full dataset from the Fall 2018 semester to the Spring 2025 semester for all analyses involving part 2.

B. Survey method and participants

Toward the end of each semester, all students enrolled in physics major courses receive a link to the survey from one of the authors (MD), who is not a member of the Physics Department. Students complete the survey online and are given two prompts via email, typically spaced about two weeks apart. This administration procedure has been implemented consistently at the end of every semester since Fall 2018. The Institutional Review Board reviewed and approved the survey procedures prior to data collection and analysis.

After each administration of the survey, MD removes all potentially identifying information (name and student ID number) and assigns each student a random research ID. This hides identifying information from researchers in the Physics Department and also allows responses to be linked across semesters without identifying individual students. Students receive no incentives for participating in the survey (e.g., monetary compensation or academic credit), and they are informed that participation is voluntary and may be withdrawn at any time. Students are also made aware that their participation status will not be communicated to any member of the Physics Department. Given that participation is voluntary and involves no incentives, some degree of self-selection bias may occur, which should be considered when interpreting the results.

Part 1 of the survey contains the nine items listed in Table II. For the analyses of part 1, we include only surveys completed between the Fall 2020 and Spring 2024 semesters, during which the wording of all nine items remained consistent. Participants who did not complete all nine items in part 1 were removed from the dataset, though it is worth noting that an incomplete part 1 survey does not automatically remove the surveys from part 2 analyses. Occasionally, some students will complete the survey multiple times per semester; in such cases, only the last complete survey is retained. These criteria led to the removal of 29 surveys between the Fall 2020 and Spring 2024 semesters.

A Mahalanobis distance test (MDT) was used to screen for multivariate outliers, resulting in 13 responses flagged at the $p \leq 0.001$ significance level [29]. Upon inspection, we determined these responses did not warrant removal. The survey is administered electronically with response options constrained to only five options, so there is no opportunity for students to respond outside of these constraints. Additionally, excluding these cases would disproportionately remove students with particularly strong opinions and could lead to artificially lower means.

The final dataset for part 1 includes $N = 399$ surveys, with $N_{100} = 212$, $N_{200} = 145$, $N_{300} = 27$, and $N_{400} = 15$. These numerical levels generally correspond to students in their first year (100-level), second year (200-level), third year (300-level), and fourth year (400-level). However, as mentioned prior in Sec. I, students occasionally take courses out of order, and we acknowledge this can introduce noise into the data.

Part 2 of the survey contains 22 items—11 in part 2a and then the same 11 repeated in part 2b—listed in Table III. Because the wording of these items has remained consistent throughout all administrations, we include all surveys completed between the Fall 2018 and Spring 2025 semesters for analyses of part 2 data. We apply the same selection rules used for part 1: surveys for part 2 must be complete, and when a student submits multiple surveys within the same semester, only the last complete response is retained. These criteria resulted in the removal of 110 surveys from the dataset.

Similarly to part 1, an MDT was used to screen for potential multivariate outliers, flagging 16 responses at the $p \leq 0.001$ level. Upon inspection, none of these cases raised concern, nor did any of the 16 overlap with those flagged in part 1. As with part 1, the removal of these responses lacked sufficient theoretical justification and risked excluding students with particularly strong feelings about their own self-efficacy. The final dataset for part 2 includes $N = 565$ surveys, with $N_{100} = 267$, $N_{200} = 226$, $N_{300} = 44$, and $N_{400} = 28$.

A goal of this study is to examine course-specific changes in students' attitudes and self-efficacy regarding computational methods. A seemingly straightforward approach to this would be to assign surveys into bins corresponding to the course in which the student is enrolled. However, students often take multiple courses in a semester, which is especially true at the 300- and 400-level, but they are asked to only complete the survey once. Assigning a single survey to multiple course bins would duplicate identical responses, artificially inflating the sample size and rendering comparison between courses statistically inappropriate. To avoid this issue, we have chosen to classify surveys by the highest semester-level course in which a student was enrolled (e.g., a student taking courses 300, 342, and 330 would be categorized as "Spring, Year 3"). While this approach loses fine-grained course-level comparisons—particularly between courses offered in the same semester—it preserves the independence of observations while still allowing for broad examination of course-level patterns in the analysis. Table IV presents the sample sizes for each semester level with shorthand labels (e.g., F-Yr3 denotes students whose highest enrolled course corresponds to Fall of year 3). There is no fall semester of year 1 (F-Yr1) in the program because the current recommended sequence begins in spring of year 1 (S-Yr1), as students are expected to complete other requirements (e.g., first semester calculus) before beginning physics courses. The sample size for the final semester in the program (S-Yr4) is small for both parts of the survey, which is expected given that students at this stage are typically enrolled in a limited number of advanced courses.

TABLE IV. Sample sizes (N) of unique surveys in each semester-level category. N_1 corresponds to part 1 of the survey (Fall 2020–Spring 2024), and N_2 corresponds to part 2 (Fall 2018–Spring 2025). Semester-level categories are determined by the highest-level course in which each student was enrolled.

Semester (year)	Label	Course(s) offered	N_1	N_2
Spring (year 1)	S-Yr1	152	212	267
Fall (year 2)	F-Yr2	251, 299	145	226
Spring (year 2)	S-Yr2	300, 342	8	14
Fall (year 3)	F-Yr3	310	12	15
Spring (year 3)	S-Yr3	330, 353	7	15
Fall (Year 4)	F-Yr4	400, 401, 442	12	21
Spring (year 4)	S-Yr4	418, 490	3	7
Total			399	565

C. Measures of growth

As students progress through the undergraduate program, it is reasonable to expect changes in their self-reported attitudes and self-efficacy regarding computational physics. A straightforward way to characterize such changes is to compare the mean *present* ratings of students at different levels in the program, which mirrors a traditional pre-post framework. We refer to this approach as G_1, defined as

$$G_1 = \bar{P}_{Y00} - \bar{P}_{X00}, \tag{1}$$

where $\bar{P}_{Y00}$ and $\bar{P}_{X00}$ are the mean present ratings for students at the $Y00$ and $X00$ levels (e.g., 200- and 100-level), respectively. This method can be interpreted as the mean "growth" between the two levels. This definition assumes $Y00 > X00$ and may be applied to any suitably ordered pair of levels, e.g., 200/100, 300/100, 400/200. Since G_1 relies solely on *present* ratings, it can be applied to both part 1—whose items measure present agreement only—and part 2, which includes present ratings in addition to initial ones.

While G_1 provides a between-level comparison of *present* ratings and is the only metric by which we can estimate growth for items in part 1, it ignores the *initial* responses provided in part 2. Accounting for the *present* and *initial* ratings, we are able to define a measure for "growth" at the individual-student level. We denote this quantity as

$$G_{\text{ind}} = P - I, \tag{2}$$

where P is the student's present rating, and I is the reported initial rating from the same survey response. This individual measure establishes what "growth" in each skill means for a single student and serves as the basic unit from which additional between-level comparisons of growth in part 2 are constructed.

We can extend the individual G_{ind} metric to the group level, which allows us to examine how growth differs across groups of students. For a group of students at any given level, averaging the individual growth scores (i.e., G_{ind}) yields a single measure of how much growth that group reports overall for each item in part 2. A between-level measure can then be obtained by taking the difference between these group means, and we define that measure as

$$G_2 = (\bar{P}_{Y00} - \bar{I}_{Y00}) - (\bar{P}_{X00} - \bar{I}_{X00}), \tag{3}$$

where $\bar{P}_{Y00}$ and $\bar{P}_{X00}$ are the mean *present* ratings, and $\bar{I}_{Y00}$ and $\bar{I}_{X00}$ are the mean *initial* ratings by students at the $Y00$ and $X00$ level, respectively. Each term $(\bar{P} - \bar{I})$ represents the average individual growth (i.e., $\bar{G}_{\text{ind}}$) within that level group, such that G_2 is then the difference in reported growth between students at the $Y00$ and $X00$ levels. Because G_2 is dependent on both the *present* and *initial* ratings, it accounts for shifts in how students interpret their *initial* abilities (i.e., the response-shift bias), which is an effect not captured by G_1.

Regrouping Eq. (3) underscores the important role that initial ratings play in shaping the growth measure:

$$G_2 = G_1 + I_{\text{eff}}. \tag{4}$$

In Eq. (4), we have introduced I_{eff}, which captures the contribution of initial ratings to the difference in reported growth between groups of students. We define it as

$$I_{\text{eff}} = \bar{I}_{X00} - \bar{I}_{Y00}. \tag{5}$$

If students do not revise their understanding of their initial competencies as they progress through the program, we would find $I_{\text{eff}} \approx 0$. However, our prior work [28] shows that I_{eff} is consistently positive. Conceptually, this pattern reflects the response-shift bias described earlier, in which students revise their understanding of their initial competencies after gaining experience.

D. Efforts to establish reliability and validity

We have previously reported on earlier efforts to establish the validity and reliability of the survey in [14,27]. We examined the internal structure validity (the degree to which relationships among survey items reflect the intended underlying constructs) of the instrument using exploratory factor analyses (EFA) in SPSS (version 31.0.0.0). For part 1, we replicated the two-factor structure identified in earlier work [27], but with an updated dataset that includes the 37 surveys collected in Spring 2024. These responses represent roughly 10% of the total part 1 dataset, providing sufficient motivation to revisit the factor structure. The results of this analysis are presented in Sec. III.

To evaluate the internal structure of part 2, we conducted a separate EFA on the 22 self-efficacy items. Because students in our program have widely varying experience with "Other Programming Language," we excluded Q4 from analyses where this variation would obscure the underlying structure (e.g., curriculum-impact analyses). Undergraduate students in our program use MATLAB as the primary programming language, but we recognize that many of our students have experience with other languages as well (e.g., Python), so we include Q4 for informational purposes to capture prior or external programming experience rather than for evaluation of our curricular initiative. Full results for part 2 appear in Sec. III.

Finally, we assessed the internal consistency (a form of reliability evidence that demonstrates the extent of agreement between items) of the instrument using Cronbach's alpha. For part 1, we computed alpha coefficients for each of the two factors identified in the EFA; for part 2, we calculated alpha values for the full set of self-efficacy items. These updated reliability estimates build on earlier split-half analyses and are reported in Sec. III.

III. RESULTS

A. Preliminary analyses

Prior to conducting the following analyses, we evaluated assumptions of univariate normality, linearity, factorability, and multicollinearity for both parts of the survey. With the exception of Q2 (LaTeX), which exhibited positive skewness, all items satisfied recommended thresholds for skewness,

kurtosis, linearity, and interitem correlations. A robust presentation of these preliminary analyses is provided in Appendix A.

B. Exploratory factor analyses

For each EFA conducted across both parts of the survey, we evaluated the factorability of items using Bartlett's test of sphericity and the Kaiser-Meyer-Olkin (KMO) measure. The estimated number of factors to retain for each EFA conducted was guided by inspection of scree plots, Horn's parallel analysis, and minimum average partial (MAP) test. All EFAs were conducted in SPSS using principal PCA with direct oblimin set to zero (i.e., allowing the factors to correlate). Appendix B provides further details for these analyses.

Consistent with prior work [27], the part 1 EFA yielded a stable two-factor structure. Reliability was assessed using Cronbach's α for both factors and indicates strong internal consistency for each factor ($\alpha_{F_1} = 0.79$, $\alpha_{F_2} = 0.80$). Because these findings update previously reported work, all detailed results, including factor loadings, are provided in Appendix B.

The part 2 EFA was conducted separately for the present ratings (part 2a) and the initial ratings (part 2b). The preliminary analyses for both parts 2a and 2b indicated sufficient interitem correlations to justify proceeding with the factor analysis. We chose to retain two factors for each part based on the results of the Horn's parallel analysis (HPA), MAP test, and scree plot examination. Reliability was assessed using Cronbach's α. For part 2a, $\alpha = 0.86$ for factor 1 and $\alpha = 0.70$ for factor 2, indicating acceptable internal consistency within each factor. For part 2b, the α values were 0.89 for factor 1 and 0.79 for factor 2, again suggesting relatively strong internal consistency across both factors. Full diagnostic statistics for the part 2a and 2b preliminary analyses can be found in Appendix B. Table V presents the factor loadings, communalities, and percent variance explained for parts 2a and 2b of the survey. The two-factor structure identified in parts 2a and 2b of the survey will serve as the basis for interpreting the response-shift effects discussed in Sec. IV D.

TABLE V. Results of the principal component analysis for part 2 items, showing factor loadings and communalities (h^2) for both the *present* (part 2a) and *initial* (part 2b) skill ratings. Loadings are from the pattern matrix, reflecting the unique contribution of each item after accounting for shared variance. Percent variance explained per factor (F_1 and F_2) is also included.

	Present			Initial		
Survey item	F_1	F_2	h^2	F_1	F_2	h^2
Q1	0.64		0.35	0.66		0.40
Q2		0.50	0.26		0.74	0.48
Q3		0.78	0.57		0.82	0.68
Q4		0.77	0.56		0.77	0.67
Q5	0.64		0.53	0.74		0.56
Q6		0.69	0.73		0.65	0.69
Q7	0.67		0.69	0.73		0.66
Q8	0.63		0.62	0.71		0.67
Q9	0.67		0.47	0.72		0.58
Q10	0.88		0.68	0.87		0.70
Q11	0.80		0.64	0.85		0.67
% Variance explained	45.25	10.20		50.69	10.67	

C. Changes in students' attitudes and self-efficacy as they progress through the undergraduate program

In the following analysis, we examine how students' attitudes and self-efficacy vary across the undergraduate program. Student survey responses are grouped by the highest course level reached (100-, 200-, 300-, or 400-level). We analyze the results for part 1 of the survey, followed by the initial and present ratings for part 2, and then reported growth in the part 2 skills using Eq. (2). Together, these results address RQ1 by describing how student responses change across levels in the curriculum.

1. Part 1: Attitudes and general self-efficacy

Figure 1 shows the mean response for each part 1 item across the four undergraduate levels. For most items, the figure indicates that mean agreement with part 1 items increases across levels, with 400-level students reporting higher levels of agreement for each item compared to 100-level students.

We conducted a one-way omnibus analysis of variance (ANOVA) for each group with a follow-up Tukey honestly significant difference (HSD) test. An independent samples t test was conducted for those comparisons indicated by the Tukey HSD to likely show significant differences between group means. Complete results for all comparisons appear in Table XII in Appendix C. The effect sizes are shown as

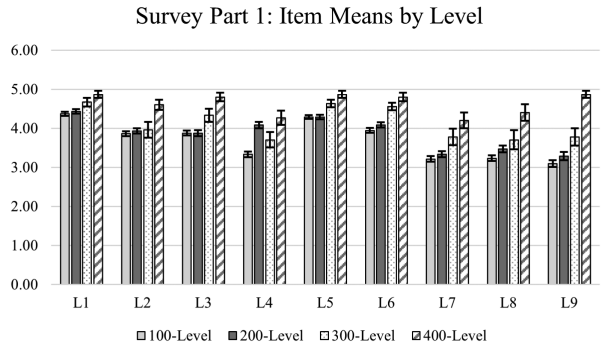


FIG. 1. The mean response for each item in part 1 of the survey, grouped by the highest course level reached by student respondents. Light gray bars represent 100-level students, dark gray bars represent 200-level students, dotted bars represent 300-level students, and striped bars represent 400-level students. Error bars indicate standard errors.

Hedges' g, which is similar to Cohen's d but is more appropriate for comparing groups of unequal sample size, and these effect sizes show that higher-level students report higher agreement for each item in part 1 than lower-level students. Levene's test for homogeneity of variance (used when comparing groups of unequal sample sizes) was conducted for each comparison, with the p values adjusted when necessary.

2. Part 2: Initial and present self-efficacy

Figure 2 displays the mean initial [panel (a)] and present [panel (b)] self-efficacy ratings for all items in part 2, excluding Q4. We have omitted Q4 from this figure and subsequent analyses because the "other programming language" is widely varied and is not directly related to evaluating the impact of our curriculum initiative. The figure shows that, for many items, present self-efficacy tends to be higher for students at advanced levels, while differences in initial ratings vary by item.

A one-way omnibus ANOVA was performed for each item in both the initial and present datasets. Tukey HSD tests identified the specific between-level comparisons for follow-up analysis, and independent samples t tests were conducted only for those comparisons. Full results appear in Table XIII in Appendix C. Effect sizes (given as Hedges' g) in Table XIII indicate that, in general, higher-level students report higher *present* self-efficacy than lower-level students for the items listed, whereas higher-level students generally report lower self-efficacy in the *initial* items than lower-level students. This pattern is reflected by the contrasting positive and negative effect sizes, where *present* items show positive effect size (e.g., $\bar{P}_{400} > \bar{P}_{100}$) and *initial* items show negative effect sizes (e.g., $\bar{I}_{400} < \bar{I}_{100}$).

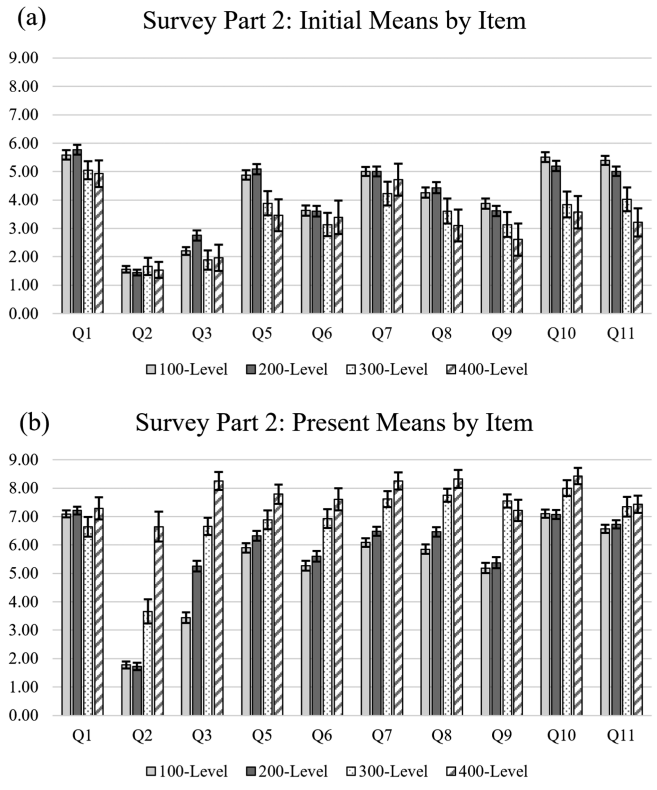


FIG. 2. The mean response for each item in part 2 of the survey, binned by highest course level reached by student respondents. Panel (a) shows the mean self-efficacy for the initial items and panel (b) shows the mean response for the present items. The light gray bar indicates students at the 100-level, the dark gray bar represents 200-level students, the dotted bar represents 300-level students, and the striped bar represents 400-level students. Error bars are the standard errors.

3. Part 2: Growth in self-efficacy

We next examined growth in students' self-efficacy for each item in part 2 (excluding Q4) using Eq. (2), $G_{\text{ind}} = P - I$, which defines growth as the difference in an individual's present and initial ratings. We then aggregated these growth scores by undergraduate level to compare mean growth among levels. A one-way omnibus ANOVA conducted on these growth scores revealed statistically significant F tests for all items except Q1 (Excel). A follow-up Tukey HSD identified specific between-level comparisons, and independent samples t tests were conducted only for those suggested comparisons. Full results appear in Table XIV in Appendix C. Table VI summarizes the comparison between 400- and 100-level students as this interlevel contrast shows statistically significant differences for all items listed. For each item listed, the effect sizes (as Hedges' g) indicate that 400-level students reported higher growth in self-efficacy than 100-level students.

Figure 3 presents the mean growth in self-efficacy for each part 2 item across levels. The figure shows that, for most items, growth tends to increase from the 100- to 400-level, which is consistent with the differences observed in the t tests.

TABLE VI. Results of the independent samples t tests, including p values (p) and Hedges' effect sizes (g) where $p < 0.05$ for reported growth (Present − Initial) in self-efficacy for all items in part 2 of the survey comparing 100- and 400-level students. The † symbol indicates p values corrected for failed Levene's tests for homogeneity of variance.

Survey item	p	g
Q2	$<0.001^{\dagger}$	3.13
Q3	<0.001	1.96
Q5	$<0.001^{\dagger}$	1.54
Q6	$<0.001^{\dagger}$	1.00
Q7	$<0.001^{\dagger}$	1.14
Q8	$<0.001^{\dagger}$	1.41
Q9	$<0.001^{\dagger}$	1.58
Q10	$<0.001^{\dagger}$	1.34
Q11	$<0.001^{\dagger}$	1.51

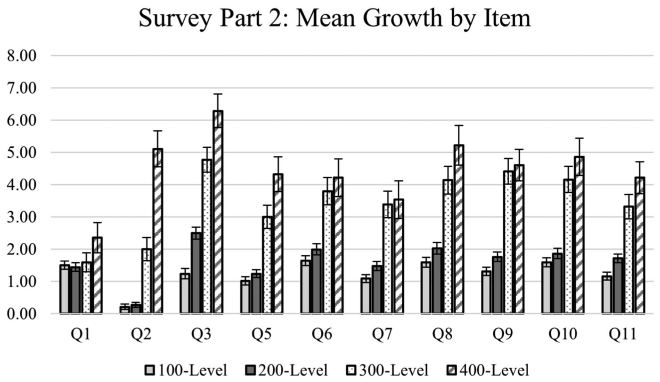


FIG. 3. The mean growth ($\bar{G}_{\text{ind}}$) for each item in part 2 of the survey, binned by highest course level reached by student respondents. The light gray bar indicates students at the 100-level, the dark gray bar represents 200-level students, the dotted bar represents 300-level students, and the striped bar represents 400-level students. Error bars are standard errors.

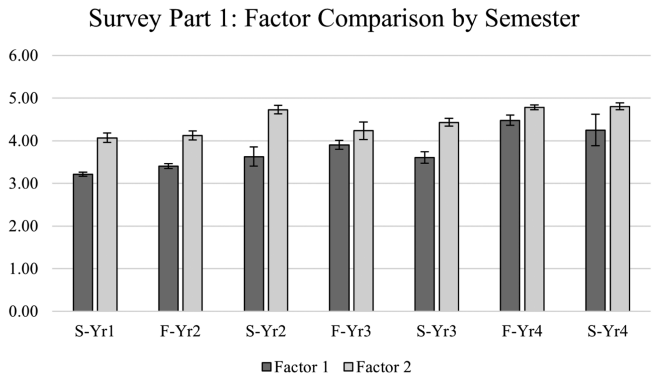


FIG. 4. The mean for each factor in part 1 of the survey per highest undergraduate semester reached. The dark gray bar represents the mean of factor 1 (items L4, L7, L8, and L9), while the light gray bar represents the mean of factor 2 (items L1, L2, L3, L5, and L6). Error bars represent the standard errors.

D. Effects of specific courses on changes in student attitudes and self-efficacy

To examine the effect of specific courses on student attitudes and self-efficacy (i.e., RQ2), we analyze how responses vary across semesters of the undergraduate program in an effort to identify where the largest changes appear.

For this analysis, we chose to bin surveys by highest semester reached in the undergraduate program (see Table IV for the semester definitions). For example, a student enrolled in courses 300, 342, and 353 during the same term would be placed in the "S-Yr3" (Spring semester of year 3) bin. This approach effectively addresses two complications with the data. First, many students will take courses out of the recommended order, which can influence how students report their attitudes and self-efficacy. If a student takes course 330 (Electricity and Magnetism) before taking course 310 (Intermediate Mechanics), that student might report higher self-efficacy with certain computational methods (e.g., solving differential equations) than what would be expected for a typical student taking courses in the recommended order. Second, students are frequently enrolled in multiple courses simultaneously, yet they complete the survey only once per semester. Assigning the same student's survey to each course in which they are enrolled would duplicate the data and inflate sample sizes. This is particularly true for upper-division students, where ≈32% of students at the 300-/400-level in the part 1 dataset are enrolled in multiple courses, and increases to ≈45% of upper-division students in the part 2 dataset. We acknowledge that some resolution is lost between individual course analysis, but binning students by highest semester reached still allows for a more granular look than simply binning students by level.

We have further chosen to complete this analysis at the factor level as opposed to the individual item level as the factor level provides a more stable measure of course-related differences across semesters. The exploratory factor analyses identified reliable two-factor structures for both part 1 and 2 of the survey, and factor scores offer a more consistent summary of these constructs. Because several semester bins contain relatively small sample sizes (particularly in the later stages of the program), item-level comparisons are significantly more subject to noise, which could obscure our ability to draw meaningful conclusions. By contrast, factor-level scores aggregate across items that measure the same underlying construct (e.g., initial and present self-efficacy), reducing item-level variability and producing comparisons that more accurately reflect the overall semester-to-semester changes within the curriculum. Given that the goal is identifying where the largest shifts occur in the program, we restrict our attention only to comparisons between adjacent semesters. Semester-level factor means (Figs. 4 and 6) were examined to identify adjacent-semester transitions associated with the largest shifts, and independent samples t tests were then conducted for those specific transitions to determine whether the observed differences were statistically reliable.

1. Part 1: Attitudinal factors

Figure 4 shows the mean response for each factor in part 1 across undergraduate semesters. Factor 1 (items L4, L7, L8, and L9) exhibits little change between adjacent semesters until a notable increase appears in F-Yr4. For factor 2 (items L1, L2, L3, L5, and L6), two substantial shifts appear: one in S-Yr2 and another in F-Yr4. A decline in the mean for factor 2 occurs between S-Yr2 and F-Yr3, followed by an increase from F-Yr3 to F-Yr4.

Independent samples t tests were conducted on visually large increases observed in Fig. 4, the results of which appear in Table VII. We chose to focus on these comparisons because our aim is to identify the semesters that contribute the most to changes in student attitudes and

TABLE VII. Results of independent samples t tests conducted between successive semesters where visually identifiable increases were observed in the means for each factor in part 1 of the survey, as well as the overall difference between students in their final fall semester (F-Yr4) and those in their first semester (S-Yr1). For both factors, we report the semester comparison (e.g., F-Yr4 vs S-Yr3), the mean difference, the effect size as Hedges' g, and the p value for the mean difference.

Factor	Semester comparison	Mean difference	g	p
Factor 1	F-Yr4/S-Yr3	0.87	1.06	0.03
	F-Yr4/S-Yr1	1.26	1.49	<0.001
Factor 2	S-Yr2/F-Yr2	0.61	1.10	<0.01
	F-Yr4/S-Yr3	0.35	0.97	0.04
	F-Yr4/S-Yr1	0.72	1.14	<0.001

self-efficacy, rather than to exhaust every possible statistically significant difference that can be found across semesters. We also include a comparison between students completing their final fall semester (F-Yr4) and those completing their first semester (S-Yr1). We use F-Yr4 for this comparison due to the substantially larger sample size relative to S-Yr4 ($N_{\text{F-Yr4}} = 21$ vs $N_{\text{S-Yr4}} = 7$).

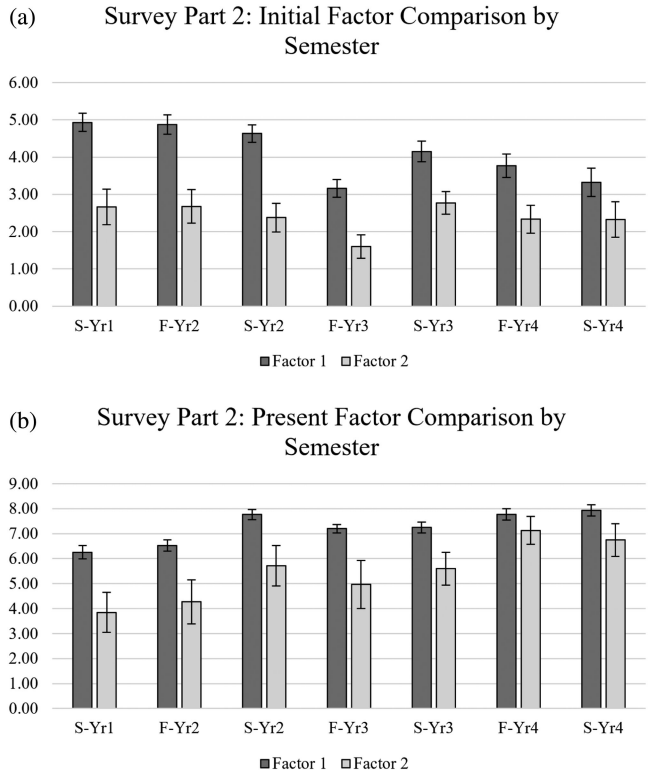


FIG. 5. The mean for each factor in parts 2a and 2b of the survey per highest undergraduate semester reached. The dark gray bar represents the mean of factor 1 (items Q1, Q5, and Q7 through Q11), while the light gray bar represents the mean of factor 2 (items Q2, Q3, Q4, and Q6). Panel (a) shows the mean response per factor for *initial* items, while panel (b) shows the mean response per factor for *present* items. Error bars represent the standard errors.

2. Part 2: Initial and present self-efficacy

Figure 5 presents the semester-by-semester mean factor scores for the initial ratings [panel (a)] and the present ratings [panel (b)], corresponding conceptually to parts 2b and 2a of the survey, respectively. Factor 1 (items Q1, Q5, and Q7 through Q11) shows a general downward trend in the initial ratings across semesters, with minimal change between adjacent semesters except for one notable shift in the fall semester of year 3. By contrast, factor 2 remains relatively stable across semesters for the initial ratings, but also shows a downward shift in the fall of year 3. For the present ratings [panel (b)], factor 1 shows a stepwise pattern with periods of relative stability punctuated by modest increases, whereas factor 2 tends to increase more steadily across semesters.

3. Part 2: Growth in self-efficacy

To examine how growth varies across semesters, we first obtained each student's G_{ind} ($G_{\text{ind}} = P - I$) for every survey item and then averaged those item-level values within each factor to produce a student's single growth score per factor. Figure 6 shows the mean of these factor-level growth scores across undergraduate semesters. Both factors exhibit their largest increases in S-Yr2, and then again in F-Yr4. Factor 1 also shows a modest rise in F-Yr3 followed by a drop in S-Yr3 before increasing once more in F-Yr4. Independent samples t tests were applied to these visually identifiable gains, and Table VIII reports the corresponding mean differences, effect sizes (as Hedges' g), and p values for each comparison. As in earlier analyses, we use F-Yr4 rather than S-Yr4 when comparing fourth-year students to those in their first semester (S-Yr1)

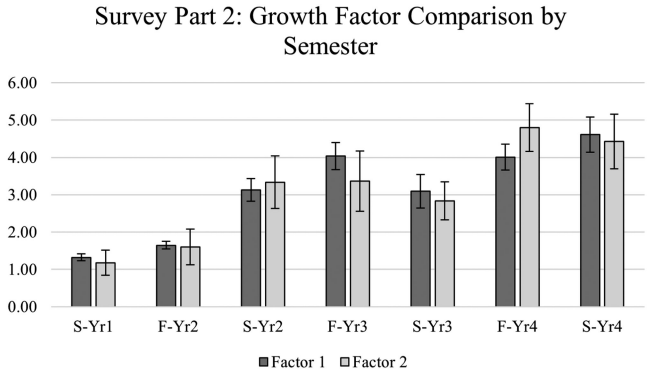


FIG. 6. The mean growth [from Eq. (2) for each individual] for each factor in part 2 of the survey per highest undergraduate semester reached. The dark gray bar represents the mean of factor 1 (items Q1, Q5, Q7 through Q11), while the light gray bar represents the mean of factor 2 (items Q2, Q3, Q4, Q6). Error bars represent the standard errors.

TABLE VIII. The results of independent samples t tests conducted between successive semesters where gains in growth were observed in the means for each factor for part 2 of the survey. For both factors we report the semester comparison (e.g., F4 and S3), the mean difference, the effect size as Hedges' g, and the p value for the mean difference. The † symbol indicates p values corrected for failed Levene's tests for homogeneity of variance.

Factor	Semester comparison	Mean difference	g	p
Factor 1	S-Yr2/F-Yr2	1.49	0.92	<0.001
	S-Yr2/S-Yr1	1.81	1.17	<0.001
	F-Yr4/S-Yr1	2.68	1.68	<0.001†
Factor 2	S-Yr2/F-Yr2	1.74	0.92	<0.01
	S-Yr2/S-Yr1	2.16	1.29	<0.001
	F-Yr4/S-Yr3	1.96	0.80	0.02
	F-Yr4/S-Yr1	3.61	2.07	<0.001†

due to the substantially larger sample size in F-Yr4 relative to S-Yr4.

Because semester-to-semester differences in growth were often small or inconsistent, we conducted simple linear regressions using highest semester reached as a predictor of students' reported growth in each factor. We applied a regression for semester-level data rather than overall level because semester ordering more directly reflects students' progression through the program. These regressions are not intended to imply linear growth but are intended instead to summarize the main effect of semester-level progression on self-efficacy. Table IX shows the regression results for each factor. For factor 1, the correlation between growth and semester was $r = 0.41$, with $R^2 = 0.17$, and the unstandardized regression coefficient was $B = 0.56$ (SE $= 0.05$, $p < 0.001$). For factor 2, the correlation was $r = 0.43$, with $R^2 = 0.18$, and the unstandardized coefficient was $B = 0.63$ (SE $= 0.06$, $p < 0.001$). These results indicate a statistically significant main effect of semester on reported growth for both factors. The correlation coefficients (r) for both factors reflect moderate correlations [29].

TABLE IX. Results from a simple regression model using the highest semester reached in the undergraduate program as a predictor for growth in students reported self-efficacy in part 2 of the survey, separated by factor. Reported are the Pearson's correlation coefficient (r), the R^2 value, unstandardized regression coefficient (B) along with the associated standard error in B (SE), and regression coefficient significance (p) for each item analyzed.

Factor	r	R^2	B (SE)	p
Factor 1	0.41	0.17	0.56 (0.05)	<0.001
Factor 2	0.43	0.18	0.63 (0.06)	<0.001

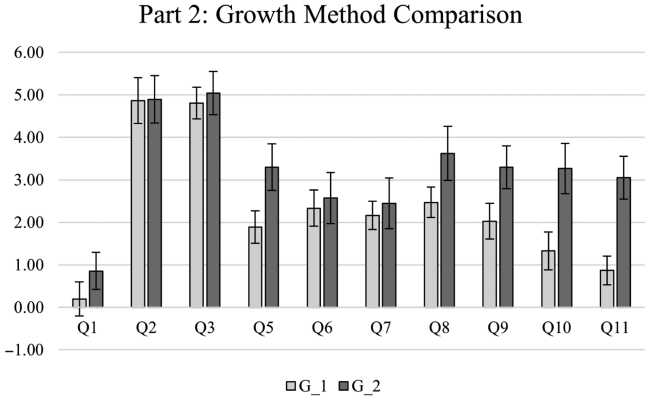


FIG. 7. A comparison of growth methods G_1 and G_2 for 100- and 400-level students. The light gray bars represent the difference between means for *present* skills (G_1), while the dark gray bars indicate G_2. Error bars are the standard error. For G_1, the standard error for items Q2 through Q9 and Q11 are corrected for failed Levene's tests for homogeneity of variance. For G_2, the standard error for items Q2 and Q5 through Q11 have been corrected for failed Levene's tests as well.

E. Quantitative comparison of methods for measuring growth in student self-efficacy

An important effect of introducing a retrospective pretest survey design is that it allows us to directly compare different methods for quantifying student growth. Figure 7 presents a comparison between growth metrics G_1 (differences in present ratings) and G_2 (differences in reported *growth*, incorporating I_{eff}) for students at the 100- and 400-levels. We focus on these two groups because most items show statistically significant differences between beginning and advanced students for both present ratings and growth measures, and these levels represent the extremes of

TABLE X. The results of independent samples t test comparisons which compare growth metrics G_1 ($\bar{P}_{400} - \bar{P}_{100}$), G_2 ($\bar{G}_{400} - \bar{G}_{100}$), and I_{eff} between 100- and 400-level students. Included are p values (p) and Hedges' effect sizes (g) where $p < 0.05$. The † symbol indicates p values corrected for failed Levene's tests for homogeneity of variance.

	G_1, 400/100		G_2, 400/100		I_{eff}, 100/400	
Survey item	p	g	p	g	p	g
Q1	0.628	⋯	0.50	⋯	0.20	⋯
Q2	<0.001†	2.30	<0.001†	3.12	0.94	⋯
Q3	<0.001†	1.63	<0.001	1.96	0.60	⋯
Q5	<0.001†	0.72	<0.001†	1.54	0.01	−0.50
Q6	<0.001†	0.83	<0.001	1.00	0.69	⋯
Q7	<0.001†	0.92	<0.001†	1.14	0.58	⋯
Q8	<0.001†	0.93	<0.001	1.41	0.05	⋯
Q9	<0.001†	0.70	<0.001	1.58	0.03	−0.43
Q10	0.003	0.59	<0.001†	1.34	<0.001	−0.67
Q11	0.007†	0.38	<0.001†	1.51	<0.001	−0.84

experience in the program. For each item, independent samples t tests were conducted, and the results appear in Table X, which reports p values and Hedges' g for differences in present scores (G_1), growth scores (G_2), and initial scores (I_{eff}). Levene's test corrections were applied where appropriate. Together, these comparisons illustrate the quantitative differences in growth estimates produced by the two methods.

IV. DISCUSSION

A. Exploratory factor analyses

1. Part 1: Attitudes and general self-efficacy

For brevity, we have included a full discussion of the EFA results for part 1 of the survey in Appendix B. The updated results indicate a reliable, two-factor structure for items in part 1 of the survey consistent with what we have previously reported [27].

2. Part 2: Initial and present self-efficacy

Exploratory factor analyses for the present (part 2a) and initial (part 2b) self-efficacy items produced consistent two-factor solutions. With the exception of item $Q2_{\text{present (LaTeX)}}$, the items showed loadings and communalities that fall within commonly accepted ranges, suggesting that the underlying factor structure represents these items well. The low communality and loading for the item $Q2_{\text{present}}$ is not unexpected, as most students do not gain meaningful experience with LaTeX until much later in the program. The poor performance of $Q2_{\text{present}}$ likely reflects this limited exposure rather than indicating a structural problem with the factor solution.

Both EFA solutions suggest the following factor groupings:

F1: Q1, Q5, and Q7–Q11 are all exclusively associated with factor 1, with loadings ranging from 0.64 to 0.88.

F2: Q2, Q3, Q4, and Q6 are all exclusively associated with factor 2, with loadings ranging from 0.50 to 0.82.

From this two-factor structure, we have tentatively named the first factor, F1, "Self-Efficacy in Computational Methods." The second factor, F2, we have tentatively named "Self-Efficacy in Computational Languages and Coding." These labels highlight a clear distinction in how students appear to organize their computational experience: specific computational methods (e.g., visualizations, numerical integration and differentiation, solving differential equations) form one factor group, while programming- and code-oriented skills (e.g., MATLAB, LaTeX, testing code) form the other. In practice, this means the factor structure separates students' sense of how well they can apply computational techniques from how well they can use the tools needed to carry out those techniques.

A notable feature of the solution is the placement of Q1 (Excel). Although Excel might reasonably be expected to fall into the "languages and coding" group, it loaded exclusively on Factor 1 in both EFAs. A possible explanation is that students are introduced to computational work through spreadsheets in the introductory labs, where they learn to organize data and generate basic plots. These early experiences could naturally lead to students conflating proficiency in Excel with specific computational techniques such as analyzing data (Q11) or creating visualizations of results (Q7). Item Q1 also shows the weakest communalities across both parts 2a and 2b of the survey, suggesting that students may interpret its scope less consistently than the others. These findings underscore the importance of monitoring weaker items such as Q1 and Q2 as additional data are collected.

B. Changes in students' attitudes and self-efficacy as they progress through the undergraduate program

1. Part 1: Attitudes and general self-efficacy

The results presented in Sec. III C show clear differences in student attitudes and general self-efficacy as they progress through the undergraduate program. Figure 1 illustrates an overall upward trend in mean responses from 100-level to 400-level students, indicating that students at later stages of the program tend to report more favorable attitudes and self-efficacy regarding computational methods in physics.

The pattern of statistically significant differences reinforces this upward trend. All nine items showed meaningful differences between 400- and 100-level students, with effect sizes ranging from moderate to large. Several items also show differences between 400- and 200-level students, while a smaller subset shows shifts between the 300- and 100-level groups. These results indicate that gains in attitudes and general self-efficacy are not restricted to a single transition point in the curriculum, but instead students report increasingly positive views and confidence with computation as they progress through the program.

When items are considered in terms of the factors identified in part 1 of the survey, an interesting pattern emerges. Items associated with factor 1 (general computational self-efficacy) tend to show more level-to-level differences than those associated with factor 2 (attitudes toward computation). In particular, items L7 and L9 (associated with factor 1) show significant differences across more comparisons than most of the other items. Item L6, though cross-loaded between factors, also shows a broader pattern of differences. This suggests that students' general self-efficacy with computational methods may shift earlier (or more noticeably) than their general attitudes about computation in physics, though both factors still show meaningful growth by the time students reach upper-division courses.

2. Part 2: Initial and present self-efficacy

The initial competency ratings (part 2b) provide clear evidence of students' response-shift bias. Items Q5, Q10,

and Q11 all show significantly lower reported initial competencies among 400-level students compared to 100-level students. This reversal, where advanced students report being *less* competent when they began the program, aligns neatly with the idea that students "did not know what they did not know" at earlier stages in the program. Then, as they gain experience with computational skills in the curriculum, their understanding of what these specific computational tasks entail increases, leading them to revise their sense of their initial abilities.

By contrast, present competency ratings (part 2a) show consistent upward trends for several items. Eight items, including Q2, Q3, and Q5 through Q10, show significant differences between 400-level students and those at the 100- and 200-levels. These results seem to indicate that students gain substantial confidence in their computational skills as they advance through the curriculum.

A comparison of the initial and present ratings reveals some interesting nuances. Some items, such as Q6 (testing code) and Q7 (visualizations), show no significant differences in *initial* ratings but do show significant differences in *present* ratings, which suggests that students maintain relatively stable perceptions of their initial competencies at higher levels, but report increased confidence in their present competency. By contrast, item Q11 (data analysis) shows significant differences in initial ratings across multiple level comparisons, but these significant differences are not reflected in the present ratings. Further, item Q10 (numerical integration/differentiation) shows significant differences in both the present and initial ratings, though the effect size is smaller for the present ratings. These patterns seem to suggest that the impact of response-shift bias varies across items, particularly for skills that lower-level students may interpret more narrowly. For example, lower-level students may be more likely to mistake Q10 (numerical integration and differentiation) as simple calculus tasks rather than computational processes.

3. Part 2: Growth in self-efficacy

The growth metric [Eq. (2)] offers a more complete picture of how students' computational self-efficacy changes with experience in the program. Figure 3 shows that upper-level students report substantially greater growth than those at the introductory levels, which is further supported by the effect sizes presented in Table XIV. With the notable exceptions of Q1 and Q4 (which was omitted from this analysis), all items show statistically significant differences across multiple level comparisons, with the vast majority showing moderate-to-large effect sizes.

The few low effect sizes occur between the 100- and 200-level students for Q3 and Q11, which is expected given the relatively small number of courses taken between these stages. By contrast, the comparisons between lower-level and upper-level students consistently yield large effect sizes, highlighting the degree to which students develop their computational skills as they advance through stages in the curriculum.

These results reinforce the value of examining both the initial and present competency ratings. Growth scores capture aspects of students' development that cannot be inferred from present or initial ratings alone, particularly for items where response-shift bias plays a significant role.

C. Effects of specific courses on changes in student attitudes and self-efficacy

1. Part 1: Attitudes and general self-efficacy

Students' attitudinal responses toward computational physics show relatively modest semester-to-semester variation across most of the undergraduate program. However, two points of more substantial change (S-Yr2 and F-Yr4) emerge when examining factor-level patterns. These changes, which appear both in the mean responses and the comparisons across semesters, offer insight into which stages of the curriculum may be contributing most meaningfully to shifts in student attitudes and general computational self-efficacy.

For factor 1 (self-efficacy in computational physics), responses remain largely stable until a noticeable increase occurs in F-Yr4. This shift could suggest that the upper-division courses taken during this semester (Physical Optics, the Physical Optics Lab, and Quantum Mechanics) provide students with experiences that strengthen their confidence with computational methods. Although earlier semesters show small fluctuations, the most substantial change for this factor clearly aligns with students' final fall semester in the program.

By contrast, factor 2 (attitudes toward computational physics) shows meaningful change earlier in the curriculum. The first noticeable shift occurs in S-Yr2, indicating that courses taken in this semester (Mathematical Physics and Modern Physics) may play an important role in shaping how students view the value or relevance of computation in physics. A second increase in F-Yr4 complements the pattern observed for factor 1, providing additional support that the experiences in the advanced courses offered that semester may reinforce or further develop students' attitudes toward computational physics.

Together, these results highlight two semesters in which attitudinal development appears most pronounced: S-Yr2, where attitudes seem to begin shifting, and F-Yr4, where both attitudes and general self-efficacy appear to show larger gains. While the current data do not allow us to isolate the effects of individual courses, these semester-level patterns point to two specific stages in the curriculum that may be particularly impactful for how students conceptualize computation in physics.

2. Part 2: Initial and present self-efficacy

For the *initial* ratings (part 2b), clear differences are observed between the two factors. Responses for factor 1

(self-efficacy in computational methods) show a general downward trend across semesters, with relatively little change between adjacent semesters except for one noticeable shift in F-Yr3. However, responses for factor 2 (self-efficacy in computational languages and coding) remain comparatively stable across most semesters, though a noticeable downward shift can also be observed in F-Yr3 for this factor. Although both factors show a dip during the same semester, the broader downward trend for factor 1 suggests that students may revise their understanding of these skills more substantially as they progress through the program, while the relative stability for factor 2 could mean that students' interpretations of these items remain more consistent over time. Because both factors appear to show a noticeable dip in the same semester (F-Yr3), it is possible that the course Intermediate Mechanics (the only course offered that semester) may have a stronger effect for how students consider their initial state.

For the *present* ratings (part 2a), the two factors again display distinct trajectories. Factor 1 appears to exhibit a stepwise pattern, characterized by periods of relative stability punctuated by modest increases across semesters, while responses to factor 2 show a more stable increase in present ratings as students advance through semesters in the program. This contrast in present ratings between factors could suggest something similar to what was observed in the initial ratings: that gains in students' perceived self-efficacy do not increase uniformly among items, but could instead be partially dependent on the underlying construct of the skill being assessed.

The relationship between the initial and present ratings across both factors highlights an important distinction in how self-efficacy appears to evolve. For factor 1, changes are more pronounced in the *initial* ratings than the *present* ratings, whereas for factor 2, this relationship is reversed: the *present* ratings appear to change more dramatically than the *initial* ratings. What this suggests is that students may be increasingly revising their understanding of the skills required for competency in computational methods, while they may be less likely to revisit their understanding of the computational language and coding tools themselves, even as their perceived proficiency in these tools grows. This contrast seems to indicate that the extent to which students reinterpret their initial self-efficacy depends, at least in part, on the underlying construct being measured, rather than reflecting a uniform response-shift effect across all computational skills.

3. Part 2: Growth in self-efficacy

Across both factors in part 2 of the survey, the largest gains in self-efficacy growth can be observed in the S-Yr2 and F-Yr4 semesters, with comparatively smaller or inconsistent changes between adjacent semesters. These patterns indicate that increases in self-efficacy tend to occur most noticeably during particular stages of the program rather than accumulating gradually over time. Notably, these semesters were also identified as points of substantial change in both student attitudes and general self-efficacy in part 1 of the survey, reinforcing the idea that these curricular stages represent important inflection points in students' development.

Although both factors show growth concentrated in similar semesters, the nature of that growth is consistent with the factor-dependent patterns observed in the initial and present ratings. For factor 1 (self-efficacy in computational methods), growth seems to be driven by changes in how students assess their *initial* self-efficacy, while for factor 2 (self-efficacy in computational languages and coding), growth appears more closely tied to changes in students' perceived *present* skill level over time. In other words, growth seems to arise from different sources depending on the factor: growth in factor 1 appears to be driven primarily by changes in *initial* ratings, while changes in the *present* ratings appear to drive growth in factor 2.

Because semester-to-semester changes in growth are often small or uneven, we also examined the overall relationship between growth and progression through the undergraduate program using a simple linear regression. While these models do not intend to imply linear development, they provide a measure of the cumulative effect that the stage in the program has on growth in students' self-efficacy. For both factors, growth shows a significant, positive correlation with the highest semester a student has reached. These results indicate that while the gains in students' self-efficacy are concentrated more sharply in specific semesters, students' self-efficacy continues to increase overall as they move through the program.

In summary, these results suggest that the growth observed in computational self-efficacy is dependent both on the stage in the program and the underlying construct being measured. The results also suggest that the measured growth emerges from the interconnection between students revising their understanding of their own initial competencies and how they assess their present abilities as they gain experience in the curriculum.

D. Quantitative comparison of methods for measuring growth in student self-efficacy

Comparing growth metrics that do and do not account for response-shift effects (i.e., comparing G_2 and G_1, respectively) reveals that the added value of a retrospective pretest design is not uniform across all survey items. For some skills, growth estimates based solely on differences in students' reported *present* self-efficacy closely match estimates that incorporate retrospective assessments of students' *initial* state. For other items, however, these two approaches yield substantially different estimates of growth, and when they differ, the difference comes from students revising how they evaluate their *initial* skill level.

For several items in part 2 of the survey, growth metrics based only on present ratings (G_1) and on reported growth (G_2) produce similar results when comparing 100- and 400-level students (e.g., Q1, Q2, Q3, Q6, and Q7). In these cases, the effective initial difference (I_{eff}) is approximately zero, indicating that students' reported initial self-efficacy does not noticeably change as they advance through the program. For such cases, accounting for response-shift effects has little impact on the resulting growth estimates, suggesting that present-only comparisons capture most of the observed difference between beginning and advanced students.

However, for a number of items (e.g., Q5, Q8, Q9, Q10, and Q11), growth estimates that incorporate retrospective initial ratings (G_2) are substantially larger than those based on present ratings alone (G_1). These differences correspond to non-negligible values of I_{eff}, indicating that students tend to revise their reported initial self-efficacy for these items as they gain experience in the curriculum. For these items, relying only on the present ratings can lead to underestimating the magnitude of growth relative to methods that incorporate response-shift effects. In other words, the observed growth for these skills arises not only from increases in students' perceived *present* self-efficacy but also from changes in how students conceptualize their *initial* level of competence.

Importantly, these differences are not randomly distributed across items but instead align strongly with the underlying factor structure. Items associated with factor 1 (self-efficacy in computational methods) appear to exhibit substantially larger growth when response-shift effects are taken into account, whereas items associated with factor 2 (self-efficacy in computational languages and coding) show little difference between growth metrics. This pattern suggests that the importance of accounting for response-shift bias may depend on the underlying construct being measured. Skills tied to specific tools or languages may be interpreted more consistently by students over time, while more abstract, methods-based skills appear more susceptible to changes in students' internal standards as they gain experience. This trend can be explicitly seen in Fig. 8, which shows the average value of I_{eff} per factor.

Notably, two items (Q1 and Q7) do not conform to the general factor-dependent trend observed. As discussed earlier, interpretations regarding Q1 are challenging due to when and how often our students gain experience with Excel. However, Q7 (visualizations of functions and other results) loads entirely on factor 1, but appears to exhibit the same $I_{\text{eff}} \approx 0$ trend that is observed primarily for items in factor 2. This discrepancy may reflect the breadth of skills encompassed by the item; some students may interpret it as referring to procedural visualizations of functions (e.g., energy level plots for a 1D infinite well in quantum mechanics), others may interpret the question as creating visualizations for real data (i.e., akin to data analysis), and some students could interpret the item as some combination of the two. However students interpret the question, it is possible that the negligible value of I_{eff} for this item could be the result of students maintaining that interpretation throughout their experience in the undergraduate curriculum. While these results show that deviations from factor-dependent patterns in I_{eff} can occur, they underscore that the observed response-shift effects should be interpreted as general patterns at the factor level rather than uniform item-by-item behavior.

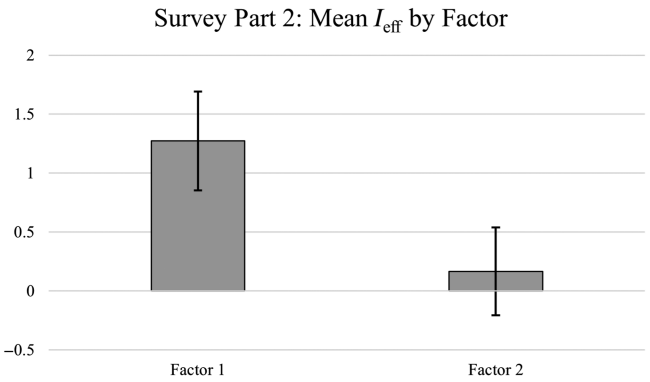


FIG. 8. The mean, absolute value of I_{eff} ($|\bar{I_{X00}} - \bar{I_{Y00}}|$) per factor in part 2 of the survey. Error bars are the standard errors. A general linear model was applied to the two groups to test the statistical significance of the difference observed. The significant F test result of $F = 9.98$ ($p = 0.002$) suggests that the difference between the two bars is statistically significant.

These results seem to indicate that growth in computational self-efficacy is not construct neutral, and that methodological choices can meaningfully shape how growth is quantified and interpreted. Retrospective pretest designs may not be universally necessary, but they appear particularly important for measuring growth in skills where students' understanding of the skill itself changes over time. For example, item Q3 (MATLAB) refers to a specific tool, so students may have a more stable sense of what "being good at using MATLAB" means over time. However, for item Q10 (numerical integration and differentiation), which is a more methods-based skill, students may be more susceptible to misinterpreting what competence with this method truly means, resulting in a greater likelihood of students "not knowing what they don't know" early in the program. More broadly, this comparison highlights that accurate measurement of growth requires attention not only to students' reported *present* abilities, but also to how students reinterpret their *initial* skill level as they gain experience in the program.

V. CONCLUSIONS

The goal of this study was to evaluate a department-wide initiative, beginning in 2018, aimed at making computational methods a routine and normalized component of undergraduate physics coursework. Using survey data

collected from undergraduate students enrolled in physics courses over several years, we examined how students' attitudes and self-efficacy with computational methods and skills evolved as they move through the physics program.

With respect to RQ1, we found that students report increasingly positive views of computation and greater self-efficacy in their computational skills as they progress through the undergraduate physics curriculum. This pattern appears in both parts of the survey: students at higher levels typically report stronger agreement with statements about the relative value of computation in physics and also report higher self-efficacy for many specific computational tasks. In practice, this means that students who are further along in the program tend to feel not only more confident in their computational abilities but also see computation as a more natural and useful component of doing physics. These results indicate that students' views and confidence regarding computation change in a consistent, positive direction as they gain experience throughout the curriculum.

To address RQ2, we found that changes in students' attitudes and self-efficacy are not evenly distributed across the undergraduate program. Instead, the largest shifts appear to be concentrated at particular stages of the curriculum, most notably during the spring semester of the second year (S-Yr2) and the fall semester of the fourth year (F-Yr4). The clustering of larger shifts during these semesters could point to stages in the curriculum where students encounter courses that emphasize different aspects of computational problem solving. For example, students in S-Yr2 commonly enroll in Mathematical Physics, which introduces them to certain computational tools and techniques (e.g., numerical integration and differentiation, matrix operations, MATLAB). Later, in F-Yr4, students commonly enroll in courses such as Quantum Mechanics and the Physical Optics Lab, which build on that foundation and place additional emphasis on skills such as visualizing functions and results, as well as data analysis. While the data do not presently allow us to attribute these shifts to individual courses, the fact that gains appear to cluster in these semesters suggests that the stage of the program in which students encounter certain computational methods may play an important role in shaping their experiences with them.

In response to RQ3, we find that measurements of growth in computational self-efficacy depend on the method by which that growth is measured. For some survey items, differences based only on the *present* self-efficacy (G_1) closely match those obtained when incorporating students' retrospective assessment of their *initial* ability (G_2). For other items, however, growth estimates that account for changes in how students assess their *initial* ability are substantially larger than those that rely on the *present* self-efficacy alone. Effectively, this means that comparisons based only on *present* ratings can underestimate growth for some computational skills, particularly when students' understanding of what those skills entail changes with experience in the program.

Importantly, the extent to which response-shift effects (i.e., the value of I_{eff}) influence measured growth is not the same for all survey items, but instead depends on the type of skill being assessed. Items associated with factor 2 (specific computational tools or languages), such as MATLAB, show relatively little response-shift effect. This means that students are not as likely to shift their understanding of their *initial* state as they are for other items, resulting in $I_{\text{eff}} \approx 0$ for items in this factor. As a consequence, similar growth patterns can be found for most of the items in factor 2 irrespective of how the growth is measured (i.e., $G_1 \approx G_2$). By contrast, most of the items associated with factor 1 (applied computational methods), such as numerical integration and differentiation, appear to show much larger differences between growth metrics, and show that the value of I_{eff} is not negligible for these cases where $G_2 > G_1$. This observed pattern for items in factor 1 challenges the fundamental assumption of traditional pre- and post-test designs—that participants do not shift their understanding of what is being measured. Notably, this construct-dependence on I_{eff} was not an intended or expected outcome of this study, but instead emerged from our efforts to compare different approaches for measuring growth in computational self-efficacy at the program level.

These patterns suggest that students' understanding of what these methods involve continues to change as they gain experience, even after they become comfortable using the underlying tools. This distinction between the growth observed with computational tools and the ability to use those tools for specific tasks reflects a difference between becoming familiar with specific procedures and developing an understanding of how those procedures are commonly applied in physics. While a small number of items deviate from this general pattern (Q1 and Q7), the factor dependence of I_{eff} suggests that retrospective pretest designs are especially relevant for skills where students' views of their *initial* competency shift over time. In practice, this means that accounting for response-shift effects may be important when measuring growth in self-efficacy for more methods-based computational skills, and that by not accounting for these effects, there is a systematic risk of underestimating the growth in students' computational self-efficacy.

This study has several limitations that should be taken into account when interpreting the results, including self-selection bias, co-enrollment of students in multiple courses, enrollment in courses out of order, and reliance on self-reported measures of ability. Future work will focus on further validating our instrument through student interviews to understand how students conceptualize the items and by examining how closely reported self-efficacy aligns with demonstrated performance on computational tasks. As additional data are collected, we hope to further clarify how specific courses contribute to the patterns of growth

observed through more fine-grained analyses. However, despite these limitations, the results of this study show that growth in computational self-efficacy for our students does not follow the same pattern for all computational skills, underscoring the importance of carefully considering both survey instrument design and the choice of growth metrics when evaluating computational self-efficacy at the program level.

ACKNOWLEDGMENTS

The authors acknowledge financial support from IU Indianapolis' STEM Education and Innovation Research Institute (SEIRI) and from National Science Foundation Grants No. DUE-2021209 and No. DUE-2439466. We also wish to thank the Partnership for Integration of Computation into Undergraduate Physics (PICUP).

DATA AVAILABILITY

The data that support the findings of this article are not publicly available because they contain sensitive personal information. The data are available from the authors upon reasonable request.

APPENDIX A: PRELIMINARY ANALYSES

Univariate normality was evaluated by examining the skewness and kurtosis of each item and comparing them with recommended guidelines suggesting acceptable skewness values fall between -2 and 2 [30] and acceptable kurtosis values fall between -7 and 7 [31]. All items in part 1 of the survey met both criteria. In part 2, item Q2 (LaTeX) fell outside the recommended skewness range for both the initial ($s = 3.495$) and present ($s = 2.073$) ratings, though its kurtosis values remained acceptable. The positive skewness reflects that LaTeX exposure typically occurs later in the curriculum. Because exploratory factor analyses are generally robust to moderate deviations from normality, and because Q2 did not distort the factor solution, it was retained in subsequent analyses.

Linearity was evaluated by examining scatterplots for all possible item pairs within each part of the survey. All pairs appear to exhibit linear relationships, providing sufficient evidence to support the assumption of linearity. Factorability was then assessed using interitem correlations via Pearson's r. For part 1, correlations ranged from $|0.20|$ to $|0.65|$, with approximately 78% showing at least moderate correlation ($\geq |0.32|$ based on recommended ranges [29]). For part 2, correlations ranged from $|0.04|$ to $|0.70|$, with about 48% reaching moderate correlation.

Multicollinearity was assessed by examining the upper range of interitem correlations. For part 1, none of the correlations exceeded the commonly recommended upper limit of $|0.70|$ [31]. For part 2, the correlation between the initial ratings for Q4 and Q6 showed a significant ($p < 0.001$) correlation of 0.70, which is just at the upper-limit threshold. Given that Q4 refers to experience with any programming language, a correlation between Q4 and the ability to test code (Q6) is conceptually consistent. Aside from this exception, no other correlation in part 2 reached the $|0.70|$ threshold, meaning that the correlations do not indicate problematic redundancy among items.

APPENDIX B: EXPLORATORY FACTOR ANALYSES

1. Survey part 1 methods and results

Prior to conducting the EFAs for each part of the survey, we evaluated whether the data met the necessary assumptions by using Bartlett's test of sphericity and the Kaiser-Meyer-Olkin (KMO) measure. Bartlett's test of sphericity assesses whether the items share enough common structure to justify extracting underlying factors while the KMO measure evaluates whether the pattern of correlations is sufficiently compact for reliable factor extraction. Together, a significant Bartlett's test combined with a KMO value above the recommended threshold of 0.70 typically indicate that the items are appropriately factorable [32]. Each EFA was conducted in SPSS using PCA with a direct oblimin rotation set to zero (i.e., allowing the factors to correlate). All coefficients below 0.32, which is the typically recommended threshold for moderate correlation, were suppressed. When determining the number of factors to retain, we relied on HPA and the minimum average partial (MAP) test, which compare the observed data to chance expectations and assess the point at which shared variance has been adequately extracted. The HPA and MAP test complement other mechanical rules of thumb (i.e., eigenvalues and examining scree plots) and are meant to account for situations in which these rules of thumb do not reliably identify the underlying factor structure by themselves [33]. Both the HPA and MAP test were carried out in STATA (version 17), as these routines are not readily available in SPSS.

For part 1 of the survey, the Bartlett's test yielded $\chi^2(36) = 1320.95$ ($p < 0.001$), and the KMO value of 0.87 exceeded the recommended threshold of 0.70, supporting the suitability of the data for factor analysis. Both the HPA and MAP tests suggested a two-factor structure for part 1, which we have chosen to retain. The PCA for part 1 explained 59.97% of the total variance, with the first factor containing four items and the second containing five. The correlation between the two factors was $r = 0.47$, indicating a moderate relationship. Communalities for items in part 1 ranged from 0.46 to 0.71, indicating that most items are well represented by the two-factor solution. Although there is some disagreement in the literature, typical guidelines suggest that communalities of 0.50 are acceptable, with values in the 0.30–0.40 range still considered adequate in some contexts [34–37]. Table XI presents the loadings, communalities, and percent variance explained for part 1.

TABLE XI. Results of the principal component analysis with factor loading and communalities (h^2) for part 1 of the survey. Factor loadings displayed are from the pattern matrix, which shows the unique contribution of an item to each factor after accounting for possible indirect correlations between an item with each factor. The percent variance explained per factor (F_1 and F_2) is also included.

Survey item	F_1	F_2	h^2
L1		0.88	0.67
L2		0.61	0.58
L3		0.57	0.56
L4	0.60		0.50
L5		0.85	0.69
L6	0.35	0.44	0.46
L7	0.81		0.71
L8	0.84		0.69
L9	0.79		0.55
% Variance explained	47.03	12.94	

As a test of internal consistency for part 1, a Cronbach's α was found for both factors identified by the EFA. Factor 1 yielded an α of 0.79, and factor 2 yielded an α of 0.80. These values generally indicate strong internal consistency within each factor. Although there is some debate about the overreliance on Cronbach's α as a sole indicator of reliability, when considered alongside the factor loadings and communalities, these results provide additional evidence for the reliability of the part 1 factor structure [38].

2. Survey part 2 methods and results

For part 2a (present skill ratings) of the survey, the Bartlett's test resulted in a value of $\chi^2(55) = 2459.46$ ($p < 0.001$), with the KMO resulting in a value of 0.89, which is above the 0.70 threshold. These results, when considered together, indicate that the interitem correlations are sufficient to justify factor analysis. The HPA and MAP tests for these items, as well as examination of scree plots, supported retaining a two-factor structure. The PCA explained 55.45% of the variance, with the first factor containing seven items and accounting for 45.25%, while the second contains four items and accounts for 10.20%. The correlation between the two factors was $r = 0.49$, indicating a moderate relationship between the factors. Communalities ranged from 0.26 to 0.73, with roughly 73% of the values exceeding 0.50, suggesting that this factor structure provides a reasonable representation of the present skill ratings.

For part 2b (initial skill ratings) of the survey, the Bartlett's test resulted in a value of $\chi^2(55) = 3029.74$ ($p < 0.001$), with the KMO resulting in a value of 0.92, which well exceeds the recommended 0.70 threshold. These results indicate that the interitem correlations are sufficient to justify factor analysis. A two-factor structure was kept based on the results of the HPA, MAP test, and scree plot examination. The PCA solution explained 61.36% of the total variance, with seven items loading on the first factor and four on the second. The correlation between factors was $r = 0.54$, which suggests a moderate correlation. Communalities ranged from 0.40 to 0.70, with most values (approximately 82%) exceeding the threshold of 0.50, suggesting that the items were reasonably well represented by the extracted factors.

Table V in Sec. III B presents the factor loadings, communalities, and percent variance explained for parts 2a and 2b of the survey.

3. Survey part 1 discussion

Exploratory factor analysis of part 1 replicated the two-factor structure previously reported [27], providing additional evidence that the attitudinal items cluster into distinct constructs. Loadings for eight of the nine items exceeded commonly recommended thresholds, and the extracted factors showed moderate correlation, supporting a structure consistent with "Self-Efficacy in Computational Physics" (F1) and "Attitudes Toward Computational Physics" (F2). Communalities were generally strong, and the total variance explained (59.97%) indicates that the two-factor model provides a stable representation of students' attitudinal responses. Only L6 exhibited notable cross loading, and L2 and L3, while associated with F2, loaded more weakly than other items within that factor. These results mirror prior findings and clarify which items would benefit most from refinement.

Informed by these results and the goal of strengthening the interpretive clarity of F2, three items (L2, L3, and L6) were revised to more clearly reflect attitudes rather than self-perceptions of ability. The revised versions remove first-person language (e.g., "I," "me") and qualify statements less ambiguously, making them more conceptually aligned with the attitudes construct. These updated items have been used in survey administrations since Fall 2024:

- L2 (revised): Physicists can use computational methods to better understand physics topics.
- L3 (revised): Computational methods can be used to understand problems that cannot be solved analytically.
- L6 (revised): Physicists can use experimental and computational methods in a complementary way to understand complex problems.

APPENDIX C: SIGNIFICANT INTERLEVEL COMPARISONS OF STUDENTS' ATTITUDES AND SELF-EFFICACY

TABLE XII. The results of the independent samples t tests, including p values (p) and Hedges' effect sizes (g) where $p < 0.05$ for items in part 1 of the survey based on the comparisons (Level Comp) suggested by the Tukey HSD *post hoc* test. The † symbol indicates p values corrected for failed Levene's tests for homogeneity of variance.

Survey item	Level Comp	p	g
L1	400/100	<0.001†	0.67
L2	400/100	0.003	0.79
	400/200	0.002	0.86
L3	400/100	<0.001	1.06
	400/200	<0.001†	1.02
L4	400/100	<0.001	0.92
	400/200	0.004	0.78
L5	400/100	<0.001†	0.86
	400/200	<0.001†	0.87
L6	300/100	0.001	0.67
	400/100	<0.001	0.93
	300/200	0.004	0.61
	400/200	<0.001	0.91
L7	300/100	0.009	0.53
	400/100	<0.001	0.95
	400/200	<0.001	0.95
L8	400/100	<0.001	1.10
	400/200	<0.001	0.90
L9	300/100	0.006	0.56
	400/100	<0.001†	1.49
	400/200	<0.001†	1.31
	400/300	<0.001†	1.12

TABLE XIII. The results of the independent samples t tests, including p values (p) and Hedges' effect sizes (g) where $p < 0.05$ for items in part 2 of the survey based on the comparisons (Level Comp) suggested by the Tukey HSD. The subscript i notation indicates questions where students were asked to provide their *initial* competency, while the subscript p notes the *present* competency. The † symbol indicates p values corrected for failed Levene's tests for homogeneity of variance. The negative effect size observed for initial items is expected and likely reflects the impact of response-shift effects on these items.

Survey item	Level Comp	p	g
$Q5_i$	300/200	0.007	−0.44
	400/100	0.011	−0.50
	400/200	0.003	−0.60
$Q10_i$	300/100	<0.001	−0.58
	300/200	0.004	−0.48
	400/100	<0.001	−0.67
	400/200	0.004	−0.58
$Q11_i$	300/100	0.001	−0.53
	400/100	<0.001	−0.84
	400/200	<0.001	−0.71

(Table continued)

TABLE XIII. *(Continued)*

Survey item	Level Comp	p	g
$Q2_p$	300/100	<0.001†	0.87
	300/200	<0.001†	0.92
	400/100	<0.001†	2.30
	400/200	<0.001†	2.42
	400/300	<0.001	1.05
$Q3_p$	200/100	<0.001†	0.62
	300/100	<0.001†	1.10
	300/200	<0.001†	0.52
	400/100	<0.001†	1.63
	400/200	<0.001†	1.11
$Q5_p$	400/100	<0.001†	0.72
	400/200	<0.001†	0.58
$Q6_p$	300/100	<0.001†	0.59
	300/200	<0.001†	0.49
	400/100	<0.001†	0.83
	400/200	<0.001†	0.73
$Q7_p$	300/100	<0.001	0.65
	300/200	0.003	0.48
	400/100	<0.001†	0.92
	400/200	<0.001†	0.76
$Q8_p$	200/100	<0.001†	0.23
	300/100	<0.001†	0.73
	300/200	<0.001†	0.54
	400/100	0.002	0.93
	400/200	<0.001†	0.77
$Q9_p$	300/100	<0.001†	0.84
	300/200	<0.001†	0.79
	400/100	<0.001†	0.70
	400/200	<0.001†	0.65
$Q10_p$	400/100	0.002	0.59
	400/200	0.003	0.59

TABLE XIV. The results of the independent samples t tests, including p values (p) and Hedges' effect sizes (g) where $p < 0.05$ for reported growth in self-efficacy (Present − Initial) for all items in part 2 of the survey based on the comparisons (Level Comp) suggested by the Tukey HSD. The † symbol indicates p values corrected for failed Levene's tests for homogeneity of variance.

Survey item	Level Comp	p	g
Q2	300/100	<0.001†	1.16
	300/200	<0.001†	1.19
	400/100	<0.001†	3.13
	400/200	<0.001†	3.26
	400/300	<0.001	1.18
Q3	200/100	<0.001†	0.47
	300/100	<0.001	1.38
	300/200	<0.001	0.82
	400/100	<0.001	1.96
	400/200	<0.001	1.36

(Table continued)

TABLE XIV. *(Continued)*

Survey item	Level Comp	p	g
Q5	300/100	<0.001†	0.94
	300/200	<0.001†	0.89
	400/100	<0.001†	1.54
	400/200	<0.001†	1.52
	400/300	0.035	0.51
Q6	300/100	<0.001	0.84
	300/200	<0.001	0.68
	400/100	<0.001†	1.00
	400/200	<0.001	0.83
Q7	300/100	<0.001†	1.07
	300/200	<0.001†	0.87
	400/100	<0.001†	1.14
	400/200	0.002†	0.93
Q8	300/100	<0.001	1.00
	300/200	<0.001	0.79
	400/100	<0.001†	1.41
	400/200	<0.001	1.18
Q9	300/100	<0.001†	1.45
	300/200	<0.001†	1.15
	400/100	<0.001†	1.58
	400/200	<0.001	1.25
Q10	300/100	<0.001	1.09
	300/200	<0.001	0.95
	400/100	<0.001†	1.34
	400/200	<0.001†	1.22
Q11	200/100	0.002†	0.28
	300/100	<0.001†	1.06
	300/200	<0.001	0.75
	400/100	<0.001†	1.51
	400/200	<0.001†	1.18

[1] J. M. Wilson and E. F. Redish, Using computers in teaching physics, Phys. Today **42**, No. 1, 34 (1989).

[2] E. F. Redish, Is the computer appropriate for teaching physics?, Comput. Phys. **7**, 613 (1993).

[3] T. J. Atherton, Resource letter CP-3: Computational physics, Am. J. Phys. **91**, 7 (2023).

[4] R. H. Landau, Resource letter CP-2: Computational physics, Am. J. Phys. **76**, 296 (2008).

[5] E. Behringer *et al.*, *AAPT Recommendations for Computational Physics in the Undergraduate Physics Curriculum* (American Association of Physics Teachers, College Park, MD, 2016), https://www.aapt.org/Resources/upload/AAPT_UCTF_CompPhysReport_final_B.pdf [accessed July 1, 2025].

[6] P. Heron and L. McNeil, *PHYS21: Preparing Physics Students for 21st Century Careers* (American Physical Society, College Park, MD, 2016).

[7] American Physical Society and American Association of Physics Teachers, *Computational Skills. EP3 Guide: Effective Practices for Physics Programs* (2021), https://ep3guide.org/guide/computational-skills/.

[8] Partnership for the integration of computation into undergraduate physics, Available at https://www.compadre.org/picup/ [accessed July 9, 2025]

[9] D. H. McIntyre, J. Tate, and C. A. Manogue, Integrating computational activities into the upper-level paradigms in physics curriculum at Oregon State University, Am. J. Phys. **76**, 340 (2008).

[10] D. M. Cook, Computation in the Lawrence physics curriculum, Available at https://www2.lawrence.edu/dept/physics/ccli/compatlu.pdf (2006) [accessed July 11, 2025]
[11] PhysPort, Available at https://www.physport.org/ [accessed July 9, 2025]
[12] M. D. Caballero, M. A. Kohlmyer, and M. F. Schatz, Implementing and assessing computational modeling in introductory mechanics, Phys. Rev. Phys. Educ. Res. **8**, 020106 (2012).
[13] H. C. Sabo, T. O. B. Odden, and M. D. Caballero, How do we assess computation in physics?, arXiv:2308.15983.
[14] A. Gavrin, G. Vemuri, and D. Maric, Preliminary efforts to evaluate an initiative introducing computation across the undergraduate physics curriculum., *presented at PER Conf. 2021, Virtual Conference*, 10.1119/perc.2021.pr.Gavrin.
[15] I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, Am. J. Phys. **53**, 1043 (1985).
[16] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[17] National Research Council, *Adapting to a Changing World: Challenges and Opportunities in Undergraduate Physics Education* (The National Academies Press, Washington, DC, 2013).
[18] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. Phys. Educ. Res. **10**, 020119 (2014).
[19] L. Bao and K. Koenig, Physics education research for 21st century learning, Discip. Interdiscip. Sci. Educ. Res. **1**, 2 (2019).
[20] L. J. Cronbach and L. Furby, How should we measure "change"—or should we?, Psychol. Bull. **74**, 68 (1970).
[21] G. S. Howard and P. R. Dailey, Response-shift bias: A problem in evaluating interventions with pre/post self-report measures, Eval. Rev. **4**, 93 (1980).
[22] G. S. Howard, K. M. Ralph, N. A. Gulanick, S. E. Maxwell, and S. K. Gerber, Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests, Appl. Psychol. Meas. **3**, 1 (1979).
[23] J. Kruger and D. Dunning, Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments, J. Pers. Soc. Psychol. **77**, 1121 (1999).
[24] C. C. Pratt, W. M. McGuigan, and A. R. Katzev, Measuring program outcomes: Using retrospective pretest methodology, Am. J. Eval. **21**, 341 (2000).
[25] S. V. Chasteen and R. Chattergoon, Evaluating impact using post-then-pre methodology, presented at PER Conf. 2019, 10.1119/perc.2019.pr.Chasteen.
[26] K. Miller, T. Espinosa, I. Araujo, and I. Gallegos, Response-shift bias in student self-efficacy during an actively taught physics course, Phys. Rev. Phys. Educ. Res. **19**, 020167 (2023).
[27] A. Gavrin, G. Gallagher, D. Maric, and G. Vemuri, An exploratory factor analysis of a survey measuring undergraduate student attitudes of computational methods in physics, *presented at PER Conf. 2024, Boston, MA*, 10.1119/perc.2024.pr.Gavrin.
[28] G. Gallagher, G. Vemuri, A. Gavrin, and D. Maric, Evaluating students' self-efficacy with respect to computational methods in the undergraduate physics curriculum, *presented at PER Conf. 2024* .10.1119/perc.2024.pr.Gallagher
[29] L. Tabachnik, *Using Multivariate Statistics* (Pearson Education Limited, New York, NY, 2012).
[30] D. L. Bandalos and S. J. Finney, Factor analysis exploratory and confirmatory, in *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (Routledge, Boston, MA, 2018).
[31] E. Knekta, C. Runyon, and S. Eddy, One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research, CBE Life Sci. Educ. **18**, 1 (2019).
[32] M. W. Watkins, Exploratory factor analysis: A guide to best practice, J. Black Psychol. **44**, 219 (2018).
[33] B. P. O'Connor, SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's map test, Behav. Res. Methods Instrum. Comput. **32**, 396 (2000).
[34] Z. Karimian and F. Chahartangi, Development and validation of a questionnaire to measure educational agility: A psychometric assessment using exploratory factor analysis, BMC Med. Educ. **24**, 1284 (2024).
[35] S. Sovey, K. Osman, and M. E. Mohd-Matore, Exploratory and confirmatory factor analysis for disposition levels of computational thinking instrument among secondary school students, Eur. J. Educ. Res. **11**, 639 (2022).
[36] R. C. MacCallum, K. F. Widaman, S. Zhang, and S. Hong, Sample size in factor analysis, Psychol. Methods **4**, 84 (1999).
[37] B. G. Tabachnick and L. S. Fidell, *Using Multivariate Statistics*, 5th ed. (Pearson Education, Boston, MA, 2007).
[38] K. S. Taber, The use of Cronbach's alpha when developing and reporting research instruments in science education, Res. Sci. Educ. **48**, 1273 (2018).